An Advanced Technique for Preventing Pharming and Flooding Attacks on Hybrid Networks

[1]Prof. Akshay Agrawal, [2]Manas S. Mhapuskar, [3]Aishwarya S. Desai, [4]Kirti B. Ahire

[1]Asst. Professor, [2,3,4]UG Student, [1,2,3,4]Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India.

[2]manasmhapuskar2009@gmail.com, [3]ash.angel2x@gmail.com, [4]ahirraokirti36@gmail.com

Abstract: Everyone in the connected world is aware of what a network is; it is a reticulum of interconnected computers. There are large networks and small networks, but size is impertinent in terms of importance of securing a Network. The incentive of network security is essential to prevent loss, through misuse of data. There are several potential pitfalls that may come up if network security is not implemented properly. This paper focuses on two attacks namely Flooding and Pharming attacks. Pharming attack is an upgraded version of phishing attack and the main aim is to detect pharming attack using multiple server responses. This surveillance deals with both wired and wireless communication called as hybrid network. The number of pharming attacks in network is proliferating and there is a urge to prevent these attacks such that the user won't lose its own confidential information. In flooding, theare are two nodes one sender and one receiver. There is an attacker insider which corrupts the sensor node which is required to route the ACK packet from sender to receiver. The ACK packet needs to be delayed and dropped by using Network Coding and prevent the flooding attacks.

Keywords* — *N-gram, Network Coding, Multiple Acknowledgement.

I. INTRODUCTION

The Internet provides a marvellous medium for expanding the horizon of communication and business. Since all the statistics sent to the Internet is basically public, the demand for surveillance becomes crucial. The Internet transports an immeasurable range of data resources and facilities. The resources, hardware and software elements of Internet are the target of malignant attempts to acquire illegitimate access to cause interruptions and commit fraud. The most perilous component of security might be the ability to provide trust and confidence to transactions over the Internet. Defence framework and web monitoring has become a vital component of the computer security in order to handle these attacks. An incursion from inside of a network is very difficult to detect, because defence system will anticipate that it is a normal activity.

In pharming attack the risk is infinitesimal to the user. The fake site is so similar to the original site, which creates an illusion to the user that the visited URL is the legitimate one as the visual facet of the fake website is also indistinguishable to the genuine one. Pharming can be carried out either by changing the hosts file on a victim's computer or by exploiting vulnerability in DNS server software. Users who visit web pages trust their browser to avert malicious web sites from leveraging their machines to attack the system.

Pharming is an internet scamming technique in which malevolent code is installed on a persons' computer or server misleading users to duplicitous webpage. Pharming attacks can be avoided by making the user alert about the fake website. An advanced approach - that combines an IP address check, textual webpage content analysis, using the information provided by multiple DNS servers - with proper results that denotes its authenticity and effectiveness to diagnose pharming attacks at the client-side.

Also in network coding the most prominent attack occurs which is known as "Flooding Attack". In flooding the attacker resides between the sender node and the receiver node. Flooding causes pollution at the receiver node. Many compromised hosts are assembled to send useless packets to block a victim, or its Internet connection, or both, thus completely taking the victim off the Internet. The attacker node spoofs its IP address and uses multiple intermediate nodes to inundate other nodes with traffic.

Pharming attack is very dangerous to users who are dealing with sensitive accounts like user name, password and credit card number, since it is important information relating to personal and financial data. That's why keeping user's aware attention of forgery websites is an important issue.

In flooding the attacker resides between the sender node and the receiver node. In flooding, many compromised hosts are assembled to send useless packets to block a victim, or its Internet connection, or both, thus completely taking the victim off the Internet. The approach is for detection and prevention of flooding attack by using multipath-ACK technique.

II. LITERATURE SURVEY

A. *Detection by ICMP feedback*

The Internet Control Message Protocol (ICMP) is an assisting protocol in the internet protocol suite. It is utilized by network tools, to dispatch faulty messages and operational information showing, that a demanded facility is unavailable or that a host could not be reached. The hacker sends SYN packet to the server using some other IP address. After that the server receives SYN packet and acknowledges to the corresponding IP address. But the client doesn’t understand the SYN+ACK packet.

The implementation starts when the server replies SYN+ACK, with the ICMP messages added and it is sent to the client. The approach is used to get the information whether the client receives from SYN+ACK or not. Since, the packet is sent from the server to the client via router, it then identifies the reply from the client and ceases sending the message to the client, concluding that it is work of a hacker so the server identifies the SYN flooding attack.

B. *Detection by tracing the route*

This approach is commenced by tracing the route of the corresponding message where it is started from. The ICMP message can be fixed so that the last router to send original IP address of the router. The server can scrutinize the source IP address and original IP address whether it is accurate or not thus it identifies the SYN flooding attack. In the existing framework, there is no time computation mechanism for defence but it is executed in proposed system. This defence technique is used to locate the SYN flooding unambiguously with short time span and less SYN packets.

C. *Phishing Attack*

Phishing is a social engineering technique deployed to take benefit of human ignorance. It allows people to exploit the weaknesses in web security technologies. As a general engineering technique, it's a logical prolongation of dressing up in a hoax uniform to gain access to a certain area. Phishing is generally conducted by e-mail spoofing or prompt messaging and it often adjure users to enter personal information at a fake website whose form and feel are almost similar to the legitimate one.

D. *Domain Name System (DNS)*

The cause of DNS makes it a very subtle area; as this is the place the client connection is orientated. The probability of a black-hat succeeding in hacking DNS is enormous (a user can be directed to a host controlled by a hacker, whatever service he might be using: http, ftp, telnet). DNS server is any computer registered to join the Domain Name System.

A DNS server runs specific networking software, which has a public IP address, and holds a database of network names and addresses for other Internet hosts’, provides means to know the IP address of any host on the Internet.

E. *Naïve Bayes (NB)*

Naïve Bayes is an approach for evaluating probabilities of individual variable value where a class is provided from training data which then allows the use of these probabilities to classify new entities which is termed in Bayesian statistics dealing with a coherent probabilistic classifier by applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. A naïve Bayes classifier infers that the existence (or exclusion) of a attribute of a class is unrelated to the existence (or exclusion) of any other attribute.

F. *An O(ND) Difference Algorithm and Its Variations*

The problems of finding a longest common subsequence of two sequences A and B and a shortest edit script for transforming A into B have long been known to be dual problems**.** In this approach, they are shown to be equivalent in terms of finding the shortest path in an edit graph.

Using this perspective, a simple O(ND) time and space algorithm is developed where N is the sum of the lengths of A and B and D is the size of minimum edit script for A and B. The algorithm performs well when differences are small and is consequently fast in typical applications.

III. AIMS AND OBJECTIVES

Many approaches have been proposed to handle Pharming and Flooding attacks. These approaches address diverse aspects of these complex threats, such as attack prevention, detection or response. The client and server side’s pharming attacks are analysed, and better false negative and false positive results have been obtained. To detect the pharming website based on the heuristic techniques using this novel approach, which has more precision recall and threshold frequency values over the existing techniques.

The Pharming attack is very treacherous to users who are dealing with sensitive information like bank details. That's why making the user aware of fraud websites is a supreme issue.

In flooding the attacker resides between the sender node and the receiver node. In flooding, a large number of compromised hosts are assembled to send useless packets to block a victim, or its Internet connection, or both, thus completely taking the victim off the Internet. This approach is for detection and prevention of flooding attack by using multipath-ACK technique.

Table 1: Comparative Study.

Sr. no	Paper Title	Author's Name	Methods	Problem	Solution
1	Textual and Visual Content-Based Anti-Phishing	Haijun Zhang, Gang Liu, Tommy W. S. Chow	A Bayesian Approach	A user-specific database of word probabilities should be consulted making filtering resource intensive. .	Use support vector machine.
2	Client – Side Pharming Attacks Detection	Ibrahim S. Alfayoumi,Tawfiq.S. Barhoom	Authoritative Domain Name Servers	Authoritative queries variation from ISP to other.	Find Anomalous Response-Time Latency and Anomalous Server Identity.
3	An O(ND) Difference Algorithm and Its Variations	Eugene W. Myers	Longest Common Subsequence and shortest path method.	Longest common subsequence and shortest path need to be calculated	Instead of processing the code in partitions we simply process the code together.
4	Signatures for network coding	Denis C, Kamal Jain, & Kristin Lauther	Symmetric cryptographic methods	Each verification requires a large number of modular exponentiations.	Fake Signatures can be generated using this algorithm. Hence, computing system is used.
5	Hashing and Authentication	Hugo Krawczyk	Linear Feedback Shift register method (Hardware based)	Works on one-time pad system, refers mostly to hardware part of system.	Leads to less bandwidth, hence its preferable to use Symmetric Systems.
6	Routing Security in Sensor Network: HELLO Flood Attack and Defence	Md. Abdul Hamid, Md. Mamun-Or-Rashid and Choong Seon Hong	Probabilistic secret sharing protocol, Bidirectional Verification	Maintenance issues like message loss, failure in detecting malicious nodes and node failures	Make use of a signal strength and time threshold based AODV-HFDP (Ad-hoc On demand Distance Routing with Hello flood Detection cum Prevention) technique.

OBJECTIVE

1. To get better in sight in pharmed and legitimate website.
2. Detection of pharming and flooding attacks.
3. To develop and design a new model for protecting users from pharming attack by making use of a powerful dual approach: IP address check and website predictability.
4. Finding the accurate classifiers to predict pharmed webpages.
5. Making appropriate use of N-gram algorithm.
6. To make proper use of multi-path ACK technique efficiently.
7. To prevent flooding of packets at receiver end.

IV. PROPOSED SYSTEM

A. Multiple Path Acknowledgement

In Multiple Path Acknowledgement, each sensor in the network continuously transmits encoded packets according to network coding scheme. Each encoded packet could choose several next-hop nodes by random GBR protocol.

The forwarder nodes will generate new encoded packets from the buffered packets and then forward them to next-hops in an intra-flow way. After successful decoding, the sink must acknowledge an ACK to the initiator to alert it to stop sending any more packets.

B. Network Coding

The ACK packet needs to be delayed and dropped, hence using Network Coding flooding attack is prevented. In network coding, two information flows are identified: the data flow and the acknowledgment (ACK) flow. Both flows can be targeted by an adversary with different consequences. An adversary attacking the data flow wants to affect the messages produced by different sources and decoded by the destinations.

The block diagram shown below gives the basic idea about all the three phases explained above. For insertion of attack, Attack traffic generator block is used. For Analysis, Model formation and the server with maximum connections are used and finally the Simulator ns2/ns3 and the parameters used for analysing the attacks help in preventing the system from the attacks. There are three main phases on which the system works-Insertion of Attack, Analysis of Attack and Prevention of attack. In the first phase, intentionally the two attacks that are flooding attack and pharming attack are inserted in the system. In the next phase that is the Analysis Phase, analysis of the system and checking whether the system is victim of attack or not is carried out. In the third phase that is the Prevention Phase; inserted attack is removed from the system.

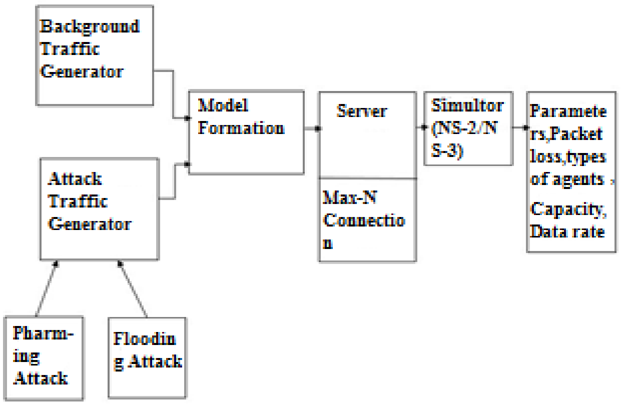


Fig 1: Proposed System

V. ALGORITHM

A. *Algorithm for Pharming attack*

Description of an N-gram-based approach to text categorization that is an indulgent of textual errors. An N-gram is an N-character slice of a longer string. Using N-grams of diverse lengths, append blanks to the commencing and terminating of the string which help in matching commencing-of-word and terminating-of-word situation. Text categorization using N-grams frequency statics: Using zipf's law[12].

Zipf's law is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency).

Let consider that

- *N* be the number of elements;
- *k* be their rank;
- *s* be the value of the exponent characterizing the distribution.

Zipf's law then calculates that, out of *N* elements, the frequency of elements of rank *k*,is given by;

$$f(k;s,N)=\frac{1/k}{\sum_{n-1}^{N}(\frac{1}{n^2})}$$

A. Generating N-Gram frequency profiles:

1. Split the text into separate tokens consisting only of letters and apostrophes. Digits and punctuation are discarded.

2. Scan down each token, generating all possible N-grams, for N=1 to 5. Use positions that span the padding blanks, as well.

3. Hash into a table to find the counter for the N-gram, and increment it. The hash table uses a conventional collision handling mechanism to ensure that each N-gram gets its own counter.

4. When done, output all N-grams and their counts sort those counts into reverse order by the number of occurrences. Keep just the N- grams themselves, which are now in reverse order of frequency.

5. The resulting file is then an N-gram frequency profile for the document. When the frequencies are plotted in this profile by rank, a Zipfian distribution graph is obtained.

B. Comparing and ranking frequency profiles

A "Measure Profile Distance" is calculated by taking two N-gram profiles and then a simple rank-order statistic is formed.

C. Types of N-gram based characterisation

(1) Language Classification:

The classification procedure is as follows:

1. Obtained training sets for each language to be classified which is in the range of 20k to120k bytes in length.

2. Computed N-gram frequency profiles on training sets, resulting profile on the order of4k in length.

3. Compute overall distance measure between sample profiles and category profile, using out-of-place measure and pick the smallest distance for the N-gram frequency profile which is in the range of 20k to 120k bytes.

(2) Subject Classification:

The classification procedure is as follows:
Obtained training sets from webpage source code:

1. Compute N-gram frequencies on the source code.

2. Compute N-gram algorithm using the averaged 2K length.

3. Compute an overall distance measure between the source code profile with the smallest distance measure from the source code profile.

B. *Algorithm for flooding attack:*

The algorithm is as follows:

Let assume a network depicted as a graph with a source node and some destination nodes.

Consider there are two nodes, first node is of the sender who wants to send the message it is known as the source node and the second node is of the receiver which receives the message send by the sender it is known as the destination node.

The source node wants to send message to the destination node, let the message to be send be denoted by *D*.

In first step, the sender node first encodes the message *D into* a certain number of m_j packets.

Let symbolize$D$= (d_1, d_2,···, *dn*) a message of *kn* bits considered as a vector of *n* fragments $d_i \in \mathbb{F}_2 k$,$i \in [1,n]$.
The encoded packets $M_j=h_j \| p_j$ transmitted by the source and the relaying nodes in a technique using random linear network coding consists in a header *hj and* a payload p_j.

$$pj=\sum_{1}^{n} ai,jdi$$

where the coefficients $a_{i,j}$ are chosen randomly over $\mathbb{F}_q$ with q= 2^u the favourite choice in the literature. The header $h_j$contains all the coefficients $\alpha_{i,}$ which describe the payload:

$hj= (\alpha 1, j,\cdots, \alpha n, j)$.

Then the sender sends to r_1 of its neighbours encoded packets m_j for j=1, until the sender receives an ACK message.

Each of the intermediate nodes forwards and integrates the received packets m_j sent by the sender to r_2 of its neighbours until the packets reach the destination node that is the receiver. The receiver after having received at least n encoded packets, begins an attempt to decode the message D. When the receiver receives a enough data, it decodes the message D and sends the ACK (acknowledgment) packet through p different routes to the sender node.

These p routes are selected among all the routes received by the receiver, the routes wherein each packet m_j brings it all the intermediate nodes from the sender to the receiver.

As soon as the source node receives one ACK message it immediately then stops sending combinations of D that is original message. To implement this algorithm, Gradient Based Routing (GBR) and Multipath Acknowledgement is used. Using the results of GBR, a modified version of GBR is used that introduces the random selection of the next hop to create a multiple path protocol useful for network coding.

1. Gradient-Based Routing (GBR)

GBR is a well-known energy efficient routing protocol. It uses a natural gradient as a metric to forward the query towards source. The metric can be regarded as physical distance, hops or others. In this work, a query is forwarded based on the hop gradient in the sensor nodes. A node forwards the query to its neighbours including its information level about the queries. After a certain period, every sensor node build up a gradient table (GTable) which indicates the distance to its sink. When a source node outwards a packet, it chooses the next hop node which has the smallest gradient in GTable. Thus, each forwarder node will choose their next-hop in the same way. Finally, the path from source to sink is established ideally.

2. Random GBR

As the network coding process is only efficient if many forwarders combine/forward the encoded packets, modify the original GBR proposal from single path routing to multipath routing from the source to the sink, and make each packet to record its route along the path. The randomization process works in the following way, when a source node outwards a packet, it randomly chooses a next hop node which have a smaller gradient than itself in GTable. So, at each packet sent, the choice for the source node for the next hop is randomly made. Each forwarder node will choose its own next-hop nodes in the same manner and so on leading to generate multipath routing when several packets are sent.

3. Multipath ACK

As previously defined, each sensor in the network continuously transmits encoded packets as per network coding scheme. Each encoded packet could choose several next-hop nodes by random GBR protocol. The forwarder nodes will generate new encoded packets from the buffered packets and then forward them to next-hops in an intra-flow way. After successful decoding, the sink must send back an ACK to the source to notify it to stop sending any more packets. Using random GBR, several paths from the source to the sink can be obtained. Using the recorded paths, thus the opportunity of ACK being blocked by flooding attackers is reduced.

VI. MATHEMATICAL MODEL

N-Gram:

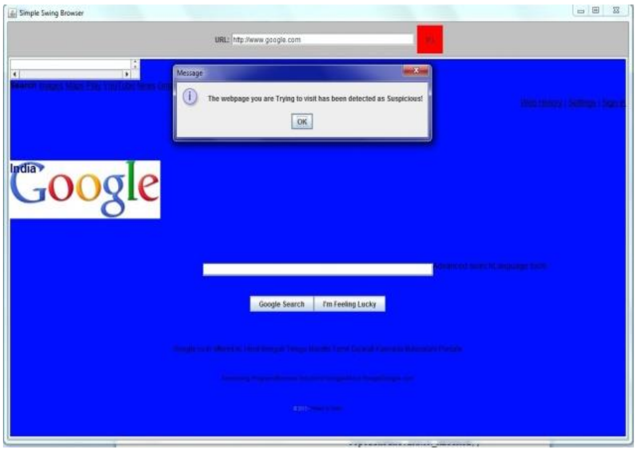
Step 1: To calculate difference between two Vectors which has two strings respectively. Hence to calculate the difference use the formula,

Difference=sqrt(sq(v1(0)-v2(0)) +sq(v1(1)-v2(1)) +sq(v1(2)-v2(2)) …+sq(v1(n-1)-v2(n-1)) (1)

Step 2: To calculate Threshold of the String. Hence, we use the formula,
Threshold = 2.486 + 0.025 * Total Length (2)
Step 3: So now, compare Difference and Threshold, if Threshold is greater than Difference, then declare that both strings are similar. If Threshold is smaller than Difference, then declare that both strings are not similar.
Step 4: Now, estimate the similarity that is normalized values between 0 and 1 also called as Probability.
Step5: Let similarity be denoted by S:
S = 4/5 + ((T-D) / (5 * T)) , if D < T
=0.8 , if T=D
=4/5 - (((D - T) * 4) / ((1 + D - T) * 5) ,if D>T, where D = Difference and T = Threshold.
Thus, Similarity is calculated using the appropriate formula as stated. The values help to check the webpage content. Intensity levels are set and then the similarity is calculated to check whether the webpage is legitimate or not.

VII. EXPECTED OUTPUT

When an URL is inserted inside the textbox of the browser and Enter key is pressed, the processing of the website starts internally. It checks the website's source code with code situated in the DNS servers. In order to check whether the website is legitimate or not, N-gram algorithm is used. This Algorithm segregates the website's source code into 'grams'. After the source code is checked, a probability value is computed by using a threshold formula. A frequency value has been predefined. If the probability value is less than it then the frequency value, then the website is considered to a malicious whereas if the probability value is greater than the frequency value, then its considered to be a Legitimate website. On the other hand, during a Flooding attack, SYN flood packet drops occur at the sender side which is avoided using Network coding



The webpage you are Trying to visit has been detected as Suspicious!

Fig 2: Preventing Pharming Attacks.

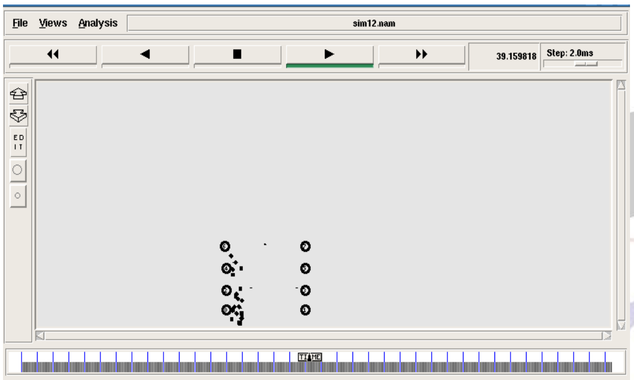


Fig 3: Preventing Flooding Attacks.

VIII. CONCLUSION

Thus, we have tried to implement the paper "Sophie Gastellier-Prevost and Maryline Laurent", "Defeating pharming attacks at the client-side", IEEE 2011 and "Yuanyuan Zhang, Wassim Znaidi, C´edric Lauradoux and Marine Minier", "Flooding attacks against network coding and countermeasures", IEEE 2011 and according to the implementation the conclusion is the composition is based on a double-step analysis and competition of multiple DNS servers - proposes an anti-pharming protection at the client-side for detecting DNS corruptions. Its administration into the client's browser can be part of a global solution that integrates protection against phishing and pharming attacks. Thus, it is implied that the IP address check is a crucial indicator of the legitimacy of a visited login website.

The purpose is to merge multiple approaches over distinct parts of the HTML source code content, to enhance the false positive rate and to regulate the processing time. Further considering the attacks against the ACK flow in network coding applications. The effect of flooding attacks is when the adversary randomly compromised the nodes. The countermeasure is grounded on multipath ACK and it is a randomized variant of GBR that enables to construct several backward paths used for the ACK sent. The choice of the routing protocol is critical and the key feature is the capacity to generate randomly many paths: greater are the paths of the ACK, higher is the probability to thwart flooding attacks. In future works, planning can be done to examine networks with other routing protocols and more cunning adversaries.

REFERNCES

[1] Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach, Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu, Senior Member, IEEE.

[2] Ibrahim S. Alfayoumi, Tawfiq S. Barhoom, Client – Side Pharming Attacks Detection using Authoritative Domain Name Servers.

[3] An O (ND) Difference Algorithm and Its Variations ,EUGENE W. MYERS, Department of Computer Science, University of Arizona, Tucson, AZ 85721, U.S.A.

[4] "APWG - Anti-Phishing working group." [Online]. Available: http:// www.apwg.org/

[5] "DNS advantage." Available: http://www.dnsadvantage.com/

[6] S. Bin, W. Qiaoyan, and L. Xiaoying, "A DNS based anti-phishing approach," in Proceedings of the 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, vol. 02. China, Wuhan: IEEE Computer Society, Apr. 2010, pp. 262–265.

[7] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in Proceedings of the 4th ACM workshop on Digital identity management. Alexandria, Viriginia, USA: ACM, Oct. 2008, pp. 51–6.

[8] "Google public DNS." Available: http://code.google.com/intl/fr/speed/public-dns/index.html

[9] "Levoyageur - banks in the world." [Online]. Available:http://www.levoyageur.net/banks.php

[10] "Routing Security in Sensor Network: HELLO Flood Attack and Defense ",Md. Abdul Hamid, Md. Mamun-Or-Rashid and Choong Seon Hong*

[11] "PhishTank." [Online]. Available: http://www.phishtan.com/

[12] Cavnar, William B. and Vayda, Alan J., "Using superimposed coding of N-gram lists for Efficient Inexact Matching", Proceedings of the Fifth USPS Advanced Technology Conference, Washington D.C., 1992.